The Astrophysical Journal, 606:763–773, 2004 May 10

REASONING FROM FOSSILS: LEARNING FROM THE LOCAL BLACK HOLE POPULATION ABOUT THE EVOLUTION OF QUASARS

Zoltán Haiman,[1] Luca Ciotti,[2,3,4] and Jeremiah P. Ostriker[2,5]

Received 2003 April 7; accepted 2004 January 20

ABSTRACT

We discuss a simple working scenario for the growth of supermassive black holes (BHs) at the center of spheroidal stellar systems. In particular, we assess the hypotheses that (1) star formation in spheroids and BH fueling are proportional to one another, and (2) the BH accretion luminosity stays near the Eddington limit during luminous quasar phases. With the aid of this simple picture, we are able to interpret many properties of the QSO luminosity function, including the puzzling steep decline of the characteristic luminosity from redshift $z \approx 2$ to $z = 0$: indeed the residual star formation in spheroidal systems is today limited to a small number of bulges, characterized by stellar velocity dispersions a factor of 2–3 smaller than those of the elliptical galaxies hosting QSOs at $z \gtrsim 2$. A simple consequence of our hypotheses is that the redshift evolution of the QSO emissivity and of the star formation history in spheroids should be roughly parallel. We find this result to be broadly consistent with our knowledge of the evolution of both the global star formation rate and the QSO emissivity, but we identify interesting discrepancies at both low and high redshifts, to which we offer tentative solutions. Our hypotheses allow us to present a robust method to derive the duty cycle of QSO activity, based on the observed QSO luminosity function and the present-day relation between the masses of supermassive BHs and those of their spheroidal host stellar systems. The duty cycle is found to be substantially less than unity, with characteristic values in the range $(3–6) \times 10^{-3}$, and we compute that the average bolometric radiative efficiency is $\epsilon \approx 0.07$. Finally, we find that the growth in mass of individual BHs at high redshift ($z \gtrsim 2$) can be dominated by mergers and is therefore not necessarily limited by accretion.

Subject headings: accretion, accretion disks — black hole physics — galaxies: active — galaxies: nuclei — quasars: general

On-line material: color figures

1. INTRODUCTION

The discovery of remarkable correlations between the masses of supermassive black holes (SMBHs) hosted at the centers of galaxies and the global properties of the parent galaxies themselves (see, e.g., Magorrian et al. 1998; Ferrarese & Merritt 2000; Gebhardt et al. 2000; Graham et al. 2001) begs for interpretation. Several groups have noted the natural link between the cosmological evolution of QSOs and the formation history of galaxies (see, e.g., Monaco, Salucci, & Danese 2000; Kauffmann & Haehnelt 2000; Granato et al. 2001; Ciotti & van Albada 2001; Cavaliere & Vittorini 2002; Menci et al. 2003 and references therein). The investigation of these interesting correlations looks promising not only to yield a better understanding of how and when galaxies formed but also to obtain information about the QSO population itself (Ciotti, Haiman, & Ostriker 2001; Yu & Tremaine 2002). For example, it may help us understand the well-known but puzzling fact that the characteristic QSO luminosity (obtained from the QSO luminosity function; see, e.g., Pei 1995; Madau, Haardt, & Rees 1999; Wyithe & Loeb 2002) drops from $z \simeq 2.5$ to $z \simeq 0$ by a factor of 35 ± 15. On the face of it, this result is surprising, since, as a result of accretion or mergers, the BHs can only grow, and more massive BHs are expected to be more luminous on average, provided that a sufficient amount of fuel is available.

Here we focus on a few specific points raised by the general remarks above: (1) What drives the evolution of the steep decline with cosmic time of the quasar luminosity density and of the characteristic quasar luminosity? (2) What is the reason for the observed relation between the cosmological evolution of the total emissivity in star-forming galaxies and that of the total emissivity of the quasar population? (3) How can one use scaling relations between the BH mass (hereafter $M_{\rm BH}$) and the host galaxy properties to determine the QSO duty cycle at redshift $z \simeq 0$?

We remark here that in this paper we *do not* attempt to physically model the complex problem of the interplay between star formation, BH growth, and QSO activity. There have been several promising suggestions in the literature for the nature of a dynamical coupling between the formation of the BH and its spheroid host, based on radiative or mechanical feedback from the SMBH on the gas supply in the bulge (Silk & Rees 1998; Haehnelt, Natarajan, & Rees 1998; Blandford 1999; King 2003). This approach is indeed very interesting and promising, especially when incorporated into population models for the evolution of both spheroids and quasars (e.g., Kauffmann & Haehnelt 2000; Volonteri, Haardt, & Madau 2003; Granato et al. 2004; Wyithe & Loeb 2003). Here we emphatically follow a different route: we explore the consequences that can be derived by empirically adopting a couple of simple and observationally well motivated working hypotheses. In other words, we do not attempt to *derive or*

[1] Department of Astronomy, Columbia University, 550 West 120th Street, New York, NY 10027.

[2] Princeton University Observatory, Peyton Hall, Princeton, NJ 08544.

[3] On leave from Dipartimento di Astronomia, Università degli Studi di Bologna, via Ranzani 1, 40127 Bologna, Italy.

[4] Also at Scuola Normale Superiore, Piazza dei Cavalieri 7, 56126 Pisa, Italy.

[5] Institute of Astronomy, University of Cambridge, Madingley Road, CB3 0HA Cambridge, UK.

explain our hypotheses, we just *use* them to derive rather interesting consequences.

The close correspondence between the observed history of star formation in spheroids and the evolution of the QSO emissivity has been noted and discussed in previous works (e.g., in Franceschini et al. 1999, who emphasize the fact that the two quantities are nearly parallel). The present paper extends the connection to a wider class of spheroidal systems and to higher redshifts and includes in the comparison some new effects (such as the fueling of BHs by the material returned from stellar winds) and updated data (e.g., in the computation of the radiative efficiency and of the duty cycle). We also focus attention on the steep decline of the characteristic quasar luminosity from redshift $z \approx 2$ to $z = 0$ and offer a tentative new empirical interpretation of this puzzling fact.

The rest of this paper is organized as follows. In § 2 we state our hypotheses and list the observational inputs required by our approach. In § 3 we illustrate the technique adopted and explore quantitatively its consequences by linking the star formation history to the QSO evolution and applying it to explain the decrease of QSO mean luminosity with decreasing redshift. Then, in § 4 we present robust estimates of the QSO duty cycle and derive the mean accretion efficiency. Finally, in § 5 we conclude by summarizing the main results and the implications of this work.

Throughout this paper we adopt the background cosmological model as determined by the *Wilkinson Microwave Anisotropy Probe* (*WMAP*; Bennett et al. 2003) experiment. This model has zero spatial curvature and is dominated by cold dark matter (CDM) and a cosmological constant (Λ), with $\Omega_m = 0.29$, $\Omega_b = 0.047$, and $\Omega_\Lambda = 0.71$, a Hubble constant $H_0 = 72$ km s^{-1}, an rms mass fluctuation within a sphere of radius 8 h^{-1} Mpc of $\sigma_8 = 0.9$, and power-law index $n = 0.99$ for the power spectrum of density fluctuations (Spergel et al. 2003). These values are consistent with their determinations by most other methods (Bridle et al. 2003; Bahcall et al. 1999).

2. BASIC ASSUMPTIONS AND MODEL INGREDIENTS

A widely accepted consequence of the so-called Magorrian relation, i.e., the (present-day) approximately linear relation between $M_{\rm BH}$ and M_S, the host spheroid stellar mass, is that the bulk of BH fueling in AGNs must be associated with star formation in the spheroidal components of their host galaxies (Monaco et al. 2000; Page et al. 2001; Granato et al. 2001, 2002; Cavaliere & Vittorini 2002). In this paper we examine the simplest possible form of this association, namely, the hypothesis that spheroid star formation and BH fueling are, at any time and in any system, proportional to one another with the proportionality constant independent of time and place.

Since most of the mass of BHs appears to have assembled within a narrow redshift interval $\Delta z \approx 1$ around $z \approx 2$ (Boyle et al. 2000; Stoughton et al. 2002), in practice this hypothesis needs to hold only during this redshift interval, in order to explain the local linear relation between BH and spheroid mass. One could argue that the energetic output from the forming central BH is the driving physical process that at the end will establish the galaxy mass (with the required proportionality). Alternatively, stars could form first, and then the BH is formed from reprocessed gas. In this case, a source of fuel for the BH growth with the required proportionality (namely, mass losses from the newly formed stars) is available in a natural way. One can imagine that both of the above scenarios lead to a linear BH versus spheroid mass relation at $z = 0$, but the strict proportionality of mass accretion rates into BHs and spheroids may not hold at all redshifts. Nevertheless, it is interesting to ask whether the simple hypothesis above is consistent with other observational data at both lower and higher redshifts, where *some* mass is still being added to both BHs and spheroids, since this test can reveal information about the physical process of the BH and spheroid mass assembly.

We make a second simple hypothesis, namely, that each BH has only two states: it is either "on" or "off;" we assume that the BH accretion luminosity always stays near the Eddington limit when the QSO is in the luminous or "on" phase and that the BH does not produce any radiation in the "off" state (e.g., because accretion is suppressed; Ciotti & Ostriker 1997). This is apparently different from other proposals in the literature that are variants of the "feast or famine" model (Small & Blandford 1992), which posit that the mean QSO volume emissivity declines toward redshift $z = 0$ as a result of, at least in part, a significant decrease in the fueling rate of individual objects (Cavaliere, Giacconi, & Menci 2000; Haiman & Menou 2000; Kauffmann & Haehnelt 2000). This is in part true, but in fact the overall "activity," i.e., the luminosity density evolution of QSOs, is the product of their characteristic number density $N_{Q^*}(z)$ and their characteristic luminosity $L_{Q^*}(z)$. We now know that all spheroidal galaxies contain massive BHs, which, when they radiate, do so at near the Eddington limit. In the limit that there is no evolution in time in their Eddington ratios, the decline can only be a preferential decline in the fraction of time that the massive objects spend in the "on" state.

We also note that Woo & Urry (2002) find Eddington ratios below unity; however, quasars appear to have typical ratios of $\sim$0.3, with only a few below 0.1 (see their Fig. 7). More importantly, our conclusions below rely only on the lack of a redshift evolution, and not the absolute value, of the Eddington ratio. Vestergaard (2004) estimated Eddington ratios in a sample of high-redshift quasars using an observed correlation between the size of the broad-line region and the luminosity of the quasar (the correlation is calibrated using reverberation mapping of lower redshift objects; e.g., Kaspi et al. 2000; Vestergaard 2002). She finds values ranging from $\approx$0.1 to $\gtrsim$1, with the $z \gtrsim 3.5$ quasars having somewhat higher $L/L_{\rm Edd}$ than the lower redshift population. In particular, Vestergaard estimates $L/L_{\rm Edd} \approx 0.3$ for two of the $z \gtrsim 6$ SDSS quasars. Given the uncertainties in these results, this is quite consistent with the assumption of near-Eddington accretion. Note further that in an extended lower redshift $0 < z < 1$ sample, Woo & Urry (2002) also find higher Eddington ratios toward $z = 1$, but this may represent a trend toward higher ratios at higher luminosity. Whether the trend is primarily with redshift or luminosity is an important question, but large scatter and selection effects presently preclude a firm answer.

The product $N_{Q^*}L_{Q^*}$ may decline as a result of a decline in fueling that shuts off AGN activity and primarily leads to a decline in N_{Q^*}. However, it is a separate question to ask what causes the surprising but well-observed decline with increasing time (Pei 1995; Boyle et al. 2000; Stoughton et al. 2002) in the characteristic luminosity $L_{Q^*}(z)$ of the observed quasars.

Coupled with the Magorrian relation, the above two hypotheses allow us to make several simple predictions, which will be described in detail in the following sections. Before we present our results, we list in detail the observational inputs required by our approach.

The *first observational input* of our analysis is the *present-day* luminosity function (LF) of spheroids, where the number of spheroids per unit volume with rest-frame *B*-band luminosities in the interval $(L_S, L_S + dL_S)$ is defined to be given by $\Phi_S(L_S)\,dL_S$. A composite LF was presented recently by Salucci et al. (1999), who considered the LF of four different types (E, S0, Sa/Sab, and Sbc/Scd) of galaxies separately and inferred the total spheroid LF by assuming that, on average, the spheroid components contribute 90%, 65%, 40%, and 10% of the light of the above galaxies, respectively. The composite spheroid LF is therefore represented by the sum of four different "Schechter law" distributions

$$\Phi_S(L_S) = \sum_{i=1}^{4} \frac{\Phi_{S*i}}{L_{S*i}} \left(\frac{L_S}{L_{S*i}}\right)^{-\alpha_i} \exp\left(-\frac{L_S}{L_{S*i}}\right), \qquad (1)$$

where $\log(\Phi_{S^*i}/\mathrm{Gpc}^{-3}) = 5.89,\ 5.95,\ 6.03,\ 6.45$, $\log(L_{S^*i}/L_\odot) = 10.18, 10.02, 10.10, 9.90$, and $\alpha_i = 0.95, 0.95, 1.0, 1.3$, for E and the bulges of S0, Sa/Sab, and Sbc/Scd galaxies, respectively. Benson, Frenk, & Sharples (2002) have recently derived the spheroid LF for a small sample of 90 bright field galaxies by decomposing the bulge and disk components, while Bernardi et al. (2003a) and Sheth et al. (2003) have computed (see also Yu & Tremaine 2002) the velocity function of early-type galaxies in the SDSS. While neither of these can serve as a substitute for the full spheroid LF to replace equation (1), this should undoubtedly be possible in the near future by decomposing a large sample of fainter late-type SDSS galaxies into their bulge and disk components.

The *second ingredient* is the quasar LF and its evolution with redshift,

$$\Phi_Q(L_Q, z) = \frac{\Phi_{Q*}/L_{Q*}(z)}{[L_Q/L_{Q*}(z)]^{\beta_l} + [L_Q/L_{Q*}(z)]^{\beta_h}}. \qquad (2)$$

The optical data in the rest-frame *B* band can be well fitted by pure luminosity evolution, with the characteristic luminosity L_{Q*} evolving with redshift as

$$L_{Q*}(z) = L_{Q*}(0)(1+z)^{\alpha_Q - 1} \frac{e^{\zeta z}(1 + e^{\xi z_*})}{e^{\xi z} + e^{\xi z_*}}. \qquad (3)$$

We adopt the fitting parameters given by Madau et al. (1999), $\beta_l = 1.64$, $\beta_h = 3.52$, $z_* = 1.9$, $\zeta = 2.58$, $\xi = 3.16$, and $\alpha_Q = 0.5$. The characteristic space density and luminosity are provided by Pei (1995) in a standard CDM cosmology with $H_0 = 50$ km s^{-1} as $\log(\Phi_{Q*}/\mathrm{Gpc}^{-3}) = 2.95$ and $\log[L_{Q*}(0)/L_\odot] = 13.03$: we adopt these values with appropriate redshift-dependent rescalings to our ΛCDM cosmology.

Finally, the *third ingredient* is the Faber-Jackson relation (Faber & Jackson 1976)

$$\frac{L_S}{10^{11}\ L_\odot} \simeq 0.62 \left(\frac{\sigma}{300\ \mathrm{km\ s^{-1}}}\right)^{4.2}, \qquad (4)$$

in the relatively more recent version of Davies et al. (1983), coupled with the $M_{\rm BH}$-σ relation (Ferrarese & Merritt 2000; Gebhardt et al. 2000; Yu & Tremaine 2002),

$$\frac{M_{\rm BH}}{10^9\ M_\odot} \simeq \left(\frac{\sigma}{300\ \mathrm{km\ s^{-1}}}\right)^4. \qquad (5)$$

Equation (4) is only approximately true, and the slope turns shallower for galaxies with velocity dispersions below[6] $\sigma \lesssim 170$ km s^{-1}. Likewise, the exponent in equation (5) is currently under debate (Ferrarese & Merritt 2000; Gebhardt et al. 2000); here we refrain from a critical assessment of the different values found in the literature and accept the slope of ∼4 as approximately the true value for both relations. Thus, to a good (but not necessarily perfect) accuracy, both the $M_{\rm BH}$-σ and the Faber-Jackson relations indicate a proportionality to the fourth power of the central velocity dispersion, implying the following linear relation:

$$\frac{M_{\rm BH}}{M_\odot} \simeq 0.016 \frac{L_S}{L_\odot}. \qquad (6)$$

When expressing equation (6) in terms of galaxy mass instead of luminosity, it is found that the implied median BH mass fraction to stellar mass is 0.13% of the mass of the bulge (Kormendy & Gebhardt 2001), which corresponds to mass-to-light ratios for spheroids of about 12 (in solar units).

3. LINKING THE STAR FORMATION HISTORY AND THE EVOLUTION OF QUASARS

As emphasized above, it is widely believed that the bulk of BH fueling in AGNs must be associated with star formation in the spheroidal components of their host galaxies. In this section we examine the hypothesis stated in § 2, namely, that spheroid star formation and BH fueling are, at any time and in any system, proportional to one another with the proportionality constant independent of time and system. Under the assumption that quasars radiate a fixed fraction ϵ of their accreted mass, an obvious consequence is that the redshift evolution of the QSO emissivity and of the star formation history in spheroids should be roughly parallel to each another. As we shall see, we find this result to be broadly consistent with our knowledge of the evolution of both the global star formation rate (SFR) and the QSO emissivity, but we identify interesting discrepancies at both low and high redshifts, to which we offer tentative solutions.

The evolution of the total UV luminosity density in stars at 1500 Å (galaxy rest frame) with redshift is given (in a standard CDM cosmology with $H_0 = 50$ km s^{-1}; Madau & Pozzetti 2000) by

$$\dot\rho_{S,\rm UV} = 7 \times 10^{26} \frac{\exp(3.5z)}{\exp(3.75z) + 20}\ \mathrm{ergs\ s^{-1}\ Hz^{-1}\ Mpc^{-3}}. \qquad (7)$$

This is related to the total star formation rate density (SFRD) as

$$\dot\rho_S = \frac{\dot\rho_{S,\rm UV}}{8 \times 10^{27}}\ M_\odot\ \mathrm{yr^{-1}\ Mpc^{-3}} \qquad (8)$$

for a Salpeter initial mass function (Madau, Pozzetti, & Dickinson 1998). Figure 1 shows (*dashed curve*) this SFRD, with an appropriate redshift-dependent rescaling to our adopted ΛCDM cosmology. This SFRD is close to that derived more directly in the recent work by Porciani & Madau (2001).

The evolution of the total rest-frame *B*-band luminosity density in quasars can be obtained from equations (2) and (3) as

$$j_{Q,B} = \int_0^\infty L \Phi_Q(L, t)\, dL. \qquad (9)$$

Under the assumption that quasars radiate a fixed fraction ϵ of their accreted mass (see discussion below), this is related

[6] For example, a fit to galaxies in the Virgo Cluster by Dressler et al. (1987) gives a mean exponent of ≃3.5, while Bernardi et al. (2003b) found instead a value of ≃4 for the exponent.

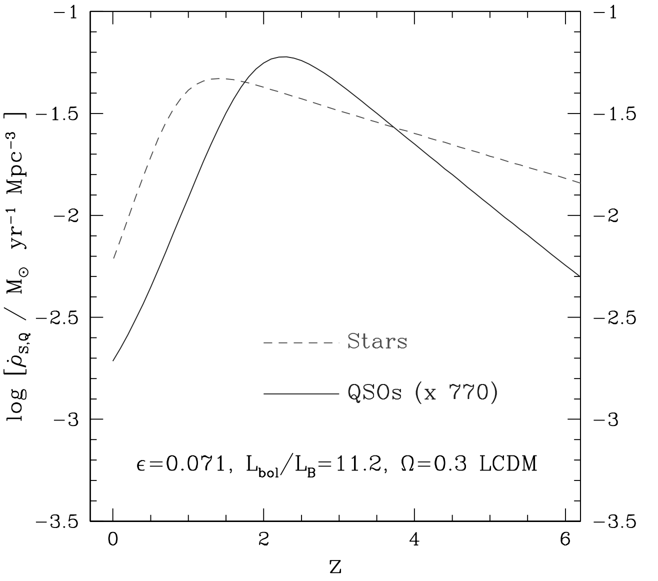

FIG. 1.—Redshift evolution of the total SFRD (*dashed curve*) and the total BARD (*solid curve*). The SFRD was adopted from Madau & Pozzetti (2000), while the BARD is obtained from the optical quasar LF, assuming a bolometric correction of $A_{\rm bol} = 11.2$ and a constant radiative efficiency of $\epsilon = 0.071$ (independent of redshift and quasar luminosity). The BARD is displaced upward by a constant factor of $1/\epsilon = 770$ for clarity of presentation. [*See the electronic edition of the Journal for a color version of this figure.*]

to the total BH accretion rate density ($\dot{\rho}_{Q,B}$; hereafter BARD) as

$$j_Q = \frac{A_{\rm bol}\epsilon c^2}{1-\epsilon}\frac{\dot{\rho}_{Q,B}}{M_\odot\ {\rm yr}^{-1}\ {\rm Mpc}^{-3}}, \qquad (10)$$

where $A_{\rm bol} = 11.2$ is the bolometric correction and $\epsilon = 0.071$ is the radiative efficiency, derived in the next section. Figure 1 shows (*solid curve*; displaced upward by a constant factor of 770 for clarity) the BH mass accretion rate density, rescaled to our adopted ΛCDM cosmology.

As is well known, both the SFRD and the BARD exhibit a steep rise from $z = 0$ to $z = 1$–2, a peak at $z \sim 1$–2, and a decline toward still higher redshifts. This is broadly consistent with their expected parallel evolution under our simple set of assumptions. Both the SFRD and the BARD still have significant observational uncertainties. While the steep decline at low redshift is relatively secure, the current SFRD and the BARD determinations could both turn out to be underestimates at high redshifts, as a result of yet undetected populations of galaxies or AGNs (e.g., as a result of dust obscuration). A critical review of the uncertainties is beyond the scope of this paper; we here simply take the current determinations at face value and examine discrepancies at both low and high redshifts from our simple model.

3.1. *Low Redshifts* ($z \lesssim 2$)

3.1.1. *Why Does the Characteristic QSO Luminosity Evolve?*

We start by pointing out the empirical fact that the bulk of BH formation, and consequently the bulk of QSO activity, must have occurred in galactic systems dominated by massive, luminous bulges. In fact, the spheroid light distribution is known to approximately satisfy a Schechter-like distribution (see, e.g., eq. [1])

$$f(L_S) \propto \left(\frac{L_S}{L_{S*}}\right)^{-\alpha} \exp\left(-\frac{L_S}{L_{S*}}\right) \qquad (11)$$

with $\alpha \approx 1.2 \pm 0.1$ (Salucci et al. 1999; Benson et al. 2002; Bernardi et al. 2003b). Then, from $M_{\rm BH} \propto L_S$, it follows that one-half of the mass in BHs is in systems with luminosity $L_S/L_{S*} \geq l_{1/2}$, where $l_{1/2}$ is defined by

$$\int_0^{l_{1/2}} l^{-\alpha+1} \exp(-l)\,dl = \frac{\Gamma(2-\alpha)}{2}. \qquad (12)$$

For $\alpha = 1.2$, this yields $l_{1/2} \simeq 0.5$. In a more detailed computation, using the composite spheroid LF given in equation (1), we find that the luminosity above which half of the integrated light is emitted corresponds to a spheroid with luminosity $M_B \approx -20.5$, only a factor of ∼2.5 fainter than the luminosity of the well-known giant elliptical M87 ($M_B = -21.42$). Thus, the bulk of the mass density of BHs resides in massive spheroidal systems, and, if the current situation is not anomalous, the bulk of the growth of SMBHs must also have occurred there (or in progenitor systems): the more common spiral galaxies have bulge luminosities typically considerably less than $L_*/2 \simeq 10^{10}\ L_\odot$. We also know that the bulk of star formation in spheroidal systems took place as early as redshift $z > 2$, as indicated, for example, by the mean stellar ages in elliptical galaxies (Hogg et al. 2002; Bernardi et al. 2003b) and of bulge populations (e.g., Proctor, Sansom, & Reid 2000; Ellis, Abraham, & Dickinson 2001), as well as the Butcher-Oemler or Gunn-Dressler effects (Margoniner et al. 2001).

At present, the disks of spiral galaxies dominate the global SFR (Fukugita, Hogan, & Peebles 1998; Benson et al. 2002; Hogg et al. 2002), and the mean age of stars in spiral systems is perhaps a factor of 2 younger than that in spheroidal systems. It follows, given our hypothesis, that BH growth in the local universe is dominated by relatively small bulges that live in galaxies denominated as spirals. Fortunately, the hypothesized relation between nuclear activity and star formation can be directly tested at low redshift. For example, Percival et al. (2001) obtained morphological information for the host galaxies of nine bright ($M_V < -25.5$) QSOs, classifying six of them as "disks" and the remaining three as "spheroids." The bulk of the local population of identified QSOs live in disk-dominated systems. The sample studied by Percival et al. (2001) was approximately 3 mag brighter than the characteristic luminosity of the local population of QSOs (Pei 1995), and almost all low-luminosity AGNs are known to reside in spiral systems. Our conclusion would therefore likely be strengthened by a survey going to magnitudes fainter than studied in the Percival et al. (2001) work, closer to L_{Q*}.

It is natural to ask what the consequences would be of the hypothesis that BH fueling, when it does happen, stays near the Eddington limit. This assumption is not unrealistic: in fact, for a handful of nearby AGNs, the BH masses can be directly estimated by reverberation mapping (or by a cruder "photoionization method"; see Wandel, Peterson, & Malkan 1999), and for these sources, the Eddington ratio can be directly inferred. From the 19 nearby AGNs listed in Table 3 and Figure 5 of Wandel et al. (1999), one derives $L = (0.01$–$0.3)L_{\rm Edd}$ for the Seyfert 1 objects, while the two QSOs have $L/L_{\rm Edd} = 0.2$ and 0.3. The bolometric correction for very hard and IR/submillimeter radiation, using the mean quasar spectrum as given by Elvis et al. (1994), is about a factor of ∼3 (with a similar value from the composite spectrum in Sazonov, Ostriker, & Sunyaev 2004), which would bring the luminosities of the two QSOs in the Wandel et al. (1999) sample close to the Eddington limit.

The assumption of always maintaining the Eddington luminosity, if applied to *an individual* BH, predicts an *increasing* QSO luminosity L_Q (in the "on" state), as a result of the trivial fact that for every BH the mass is monotonically increasing. However, this is not necessarily in conflict with observations that describe the evolution of the characteristic luminosity for a population of quasars. Clearly, if all galaxies remained equally active, then the mean luminosity inferred from the observed QSO LF would increase, but if the typical member of the population changes with time, the naive expectations may be incorrect.

Let us simply assume, as an empirically verified fact, that at the present day, star formation is active primarily in disk-dominated systems. A decrease of a factor 20–50 in the characteristic quasar luminosity L_{Q*} would then be naturally obtained by combining the Faber-Jackson ($L \propto \sigma^4$) and Magorrian ($M_{\rm BH} \propto \sigma^4$) relations with a reduction in the characteristic central velocity dispersion (σ) in the hosts of QSOs by a factor in the range 2.1–2.7. A decrease of this amount is quite natural, when one considers the mean (luminosity weighted) central velocity dispersion associated with the elliptical galaxies at redshift $z = 2$ ($\sigma \approx 400$ km s^{-1}) and that associated with spiral bulges at redshift $z = 0$ ($\sigma \approx 200$ km s^{-1}), as derived by the Faber-Jackson relation. This argument thus provides a straightforward interpretation of why the typical QSO luminosity decreases from $z = 2$ to 0. Furthermore, there is explicit observational evidence (Thomas, Maraston, & Bender 2002) that large elliptical galaxies are older than small bulges of spirals, supporting the decrease in σ toward $z = 0$ as the reason behind the decrease in the characteristic quasar luminosity.

A slightly less steep drop in the central velocity dispersion could be acceptable, by simultaneously allowing for quasar luminosities to decrease to somewhat sub-Eddington values toward low redshifts. The latter scenario is consistent with the Eddington ratios of the two quasars in the Wandel et al. (1999) sample. Applying the reverberation mapping technique to an extended quasar sample (e.g., selected from the SDSS) would distinguish directly between these two options.

3.1.2. *Why Does the Total Quasar Emissivity Evolve?*

We next consider whether the observed steep evolution of the total quasar emissivity (or BARD) is consistent with the SFRD. Figure 1 shows the BARD and SFRD. However, the SFRD includes contributions from both disk stars and spheroid stars: here we discuss corrections to this diagram to obtain the SFRD in spheroids alone.

Corrections for disk star formation are likely to become large at low redshifts. The fraction of the total stellar luminosity density at $z = 0$ contributed by stars in disks versus stars in spheroids (the latter including both the bulges of spirals and elliptical galaxies) has been estimated by several authors. Fukugita et al. (1998) and Hogg et al. (2002) find that spheroids contribute $\sim$40% of the luminosity density (but Benson et al. 2002 find that spheroids contribute significantly less than this fraction). Furthermore, it is well known that the stellar populations in present-day spheroids are old, and most models place their formation epochs at $z > 2$. Nevertheless, we are interested in the amount of ongoing star formation in these spheroids at $z = 0$. The lower limits on the ages of the spheroid stellar populations come from various methods; one of these is the colors. $B-V$ and $V-I$ magnitudes can be determined to an accuracy of $\sim$0.1 mag (e.g., Ellis et al. 2001), and these colors typically change by $\sim$1 mag in a billion years of evolution (Leitherer et al. 1999[7]). It follows that less than 10% of the present-day luminosity in these systems can arise from young stars. In turn, this implies that less than $\sim$5% of the total SFRD at $z = 0$ is occurring in spheroid systems (i.e., in the bulges of spirals). On the other hand, at redshifts of $z = 2$–4, the Lyman break galaxies are believed to be large bulge systems in the process of formation, and the fraction of the star formation seen at these redshifts associated with bulges is believed to be essentially unity.

The Butcher-Oemler (BO; also known as Dressler-Gunn) effect can also be used to derive the contribution of spheroids to the total SFR as a function of redshift in galaxy clusters. The BO effect then gives the rate of expected decline in the emission from SMBHs, since these tend to be in high-mass elliptical galaxies that are well represented in the BO clusters. Observational work on the BO effect has shown that the number fraction of blue galaxies in clusters, f_b, increases with z in a linear relation (see, e.g., Newberry, Kirshner, & Boroson 1988; Andreon & Ettori 1999; Metevier, Romer, & Ulmer 2000). For example, in one of the most detailed works, based on the analysis of 295 POSS II clusters with redshift $0 < z < 0.4$ (Margoniner et al. 2001), the authors found

$$f_b \simeq 1.3^{\pm 0.5} z + c, \tag{13}$$

where c is a small additive constant, of the order of 0.02 ± 0.01. There are important caveats in using the BO effect for the spheroid correction. First, there are large variations in f_b from cluster to cluster (e.g., the $z = 0.83$ cluster studied by van Dokkum et al. 2000 has an estimated $f_b = 0.22 \pm 0.09$). It is also not clear at the present time where the star formation responsible for this blue light is occurring. While this blue light may represent ongoing star formation in spheroids, Abraham et al. (1996) argue that the blue light arises in the disks that flare up as spirals fall into the cluster potential. However, interpreting the blue fraction f_b as the fraction of elliptical galaxies undergoing star formation, the BO effect would support the general conclusion that spheroids contribute only a few percent of the total SFR in the present-day universe, while this fraction rises steeply toward higher redshifts (but see de Propris et al. 2003 for an opposing view).

There are other promising methods to estimate the spheroid contribution to the total SFRD. For example, one could measure local starburst activity in the dense obscured centers of galaxies in the infrared bands and identify this with the local SFR in bulges. It should also be possible to measure an accurate age distribution of stellar populations in the bulges of spirals, as well as in elliptical galaxies and in large samples of SDSS galaxies, and hence to directly infer the time dependence of the SFR in spheroids as a function of redshift. An accurate measurement of the age distribution has already been achieved for a red subsample of SDSS galaxies (Jimenez et al. 2003).

In summary, we here estimate the rough fraction of the observed SFR that is associated with spheroids at each redshift by multiplying the total SFRD shown in Figure 1 by a factor $f_{\rm sph} = 0.05 + 0.95(z/2.2)^2$ at $z < 2.2$. This ensures a smooth transition from the total SFRD being dominated by elliptical galaxies at high redshift to it being dominated by disks of spirals at $z = 0$, with only residual star formation in the bulges of spirals, consistent with the arguments above. Note that our

[7] Electronic data are available at http://www.stsci.edu/science/starburst99/.

conclusions below do not rely on the explicit functional form assumed. What matters is only that the fraction of the SFR at $z = 0$ in spheroidal systems is only a few (<0.05) of the total. In Figure 2 we show (*long-dashed curve*) the corrected SFRD. As the figure demonstrates, including this correction improves the fit, in the sense of making the SFRD resemble the BARD more closely. However, intriguingly, it does appear that the decline in the spheroid formation rate from $z = 2$ to 0 is too large, by a factor of $\sim$3, when compared to the decline in the BARD. If this discrepancy holds up in future data, it would imply that the nuclear BHs can be fueled long after the star formation in the bulge has ceased. Since the bulge star formation has likely used up all the gas initially present in the bulge, the fuel would have to arrive from elsewhere.

A simple consequence of stellar evolution is that old stars that had formed in the bulge keep returning a fraction of their mass in winds. These stellar winds may provide dense, shocked material that can dissipate and serve as a fuel for the central BH. The mass-loss rate in winds in a starburst evolves as $\propto t^{-1.3}$ (where t is the time elapsed from the burst; see, e.g., Ciotti et al. 1991; Leitherer et al. 1999). We here use the wind mass-loss rate $\dot{M}_{\rm wind} = 1.5 \times 10^{-11} L_B t_{15}^{-1.36}\ M_\odot\ {\rm yr}^{-1}$ between the ages of 0.5 and 15 Gyr for a 1 $M_\odot$ model starburst galaxy, where t_{15} is the time elapsed from the burst in units of 15 Gyr and $L_B \approx 0.03$ is the B-band luminosity of the model galaxy (in units of $L_\odot$) at 15 Gyr (G. Bruzual & S. Charlot 2000, unpublished[8]). Under these assumptions, $\sim$80% of the stellar mass is eventually returned to the ambient medium.

For our purposes, we regard mass lost in winds as new material available to fuel the central BH. It is clear that the total mass return rate from a passively evolving stellar population cannot accrete onto the central BH (otherwise BH masses will be 2 or 3 orders of magnitude larger than those observed), nor can it all turn into stars (star formation in present-day elliptical galaxies is not detected at the level that would be implied). In order to solve these problems, Ciotti & Ostriker (2001) performed numerical simulations of radiative feedback modulated accretion flows onto an SMBH at the center of a "cooling flow" galaxy. They showed that only a *few percent (or less)* of the available gas lost in winds effectively accretes onto the central BH, while the accretion luminosity during short episodes of bursts stays near the Eddington value (a similar conclusion would follow in the case of mechanical feedback; e.g., Tabor & Binney 1993; Binney & Tabor 1995; Binney 1999).

Here we assume that a fraction 1.3×10^{-3} of the mass in winds accretes onto the central BH, i.e., the same fraction we had assumed for the "original" infalling gas earlier in this paper. This fits in well explicitly with the fractions inferred in Ciotti & Ostriker (2001). In Figure 2 we show (*dotted curve*) the total mass-loss rate generated in winds from spheroid stars as a function of redshift. The thick solid curve (turning into the dot-dashed curve at high redshift; see discussion below) shows the total BARD inferred from the SFRD after mass loss from winds is added. We conclude that, if a significant fraction of the wind material ends up fueling the central BH, this brings the BARD and SFRD into quite reasonable agreement (to within a factor of 2 at all redshifts).

We emphasize that our treatment in this subsection is phenomenological and complementary to theoretical semianalytical models (Haiman & Menou 2000) of the cosmological evolution of the QSO luminosity, or models based on Monte Carlo realizations of dark matter "merger trees" (see, e.g., Kauffmann & Haehnelt 2000). These models have found that to reproduce the observed decline in the QSO luminosity density, the "efficiency factor" for the fraction of gas accreted by the BH in a merger must decline toward $z = 0$. Our proposal here is radically different: instead of "starving" the BHs in each galaxy, the QSO luminosity density drops as a result of the empirically inferred drop in the SFR within spheroids and the strong bias toward smaller systems at lower redshifts. A uniform process of starvation would not lead to a decline in the typical luminosity of AGNs with time while maintaining a near Eddington output in the observed sample.

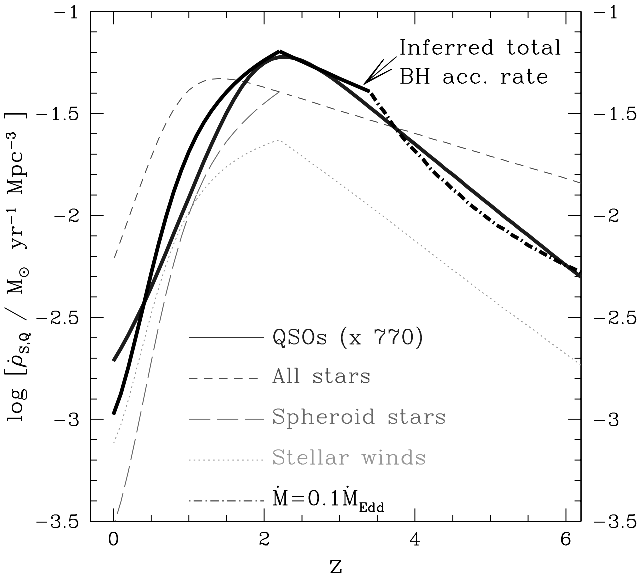

FIG. 2.—Modifications to the SFRD and BARD in Fig. 1 that would bring the two quantities to parallel each other, as required by our simple set of assumptions. The modifications include (1) a correction for the spheroid vs. total SFRD (*long-dashed curve*), (2) the possibility of fueling BHs via stellar winds (*dotted curve*), and (3) a suppression of the BARD at high redshifts due to an extra source of opacity (*dot-dashed curve*). The thick solid curve (turning into the dot-dashed curve at high redshift) shows the BARD inferred from the SFRD after these corrections are taken into account; it tracks the BARD inferred from the optical quasar LF (*thin solid curve*). [*See the electronic edition of the Journal for a color version of this figure.*]

3.2. *High Redshifts* ($z \gtrsim 2$)

As at $z \lesssim 2$, at redshifts exceeding the peak of quasar activity ($z \gtrsim 2$), the evolutions of the SFRD and BARD are, in fact, not parallel (see Fig. 1). It must be noted that the observational determinations of both quantities are much less certain at these high redshifts than at low redshifts. The presence of a population of high-redshift, dust-obscured quasars could, for example, reconcile the SFRD and BARD curves in our simple model. There is already evidence for such a population that could significantly increase the inferred BARD (e.g., Fabian & Iwasawa 1999); there is also some evidence that, unlike the optical LF, the soft X-ray quasar LF stays flat out to redshifts $z \sim 4$ (Miyaji, Hasinger, & Schmidt 2001). We here simply take the current determinations at face value and examine physical reasons that would explain the apparent discrepancy.

3.2.1. *What Steepens the Evolution of the High-z Quasar Emissivity?*

One possibility is that the fueling rate of quasars is suppressed by intrinsic physical limits to the rate of accretion. Models in which the BHs shine with approximately their

[8] The models are available at ftp://gemini.tuc.noao.edu.

Eddington luminosity can naturally explain the observed evolution of the QSO LF (Haiman & Loeb 1998; Haehnelt et al. 1998; Wyithe & Loeb 2002), by associating the rise from $z = 6$ to 2 with the increase in the nonlinear mass scale in hierarchical cosmologies. Ciotti & Ostriker (2004) suggested that at the high characteristic densities at $z \gtrsim 3$, bremsstrahlung opacity may effectively limit the mass accretion rate onto a BH to a small fraction of the usual Eddington value. This idea is attractive because it provides a physical reason for the suppression of the fueling rate and because the additional opacity may be relevant only at high redshifts, allowing "normal" accretion at $z \lesssim 3$. In fact, the data reviewed in § 3.1.1 are consistent with $L_{\rm max} = 0.1 L_{\rm Edd}$ (and a modest correction for beaming), but the effects of correspondingly increasing the Eddington time by a factor of 10 are only important at high redshifts, where it then becomes comparable to the age of the universe.

In Figure 2 we show (*dot-dashed curve*) the evolution of the emissivity for a BH, $L \propto \dot{M} \propto M \propto \exp(f_{\rm Edd} t / t_{\rm Edd})$ under the assumption that the hole grows exponentially on a timescale that is $f_{\rm Edd}^{-1} = 1/0.07 = 14$ times the Eddington time $t_{\rm Edd} = \epsilon 4.6 \times 10^8$yr. A multiplicative constant ($10^{-3.1}\ M_\odot\ {\rm yr}^{-1}\ {\rm Mpc}^{-3}$ for the curve shown in Fig. 2) can be used to represent the summed emissivity or accretion rate density of all quasar BHs, all of which are assumed to grow at the same rate from $z = \infty$. As the figure reveals, the suppression of the accretion rates in all BHs to 7% of the Eddington value would naturally result in the observed steep slope of the quasar emissivity evolution between $3 \lesssim z \lesssim 6$, while not preventing star formation from occurring in a more extended spheroid region around the BH. While attractive, this explanation suffers from a drawback, namely, the fact that if all BHs can accrete only at 10% of the Eddington rate at $z > 3.5$, then their *e*-folding time will be $\sim 5 \times 10^8$ yr, making it apparently difficult to explain how the large (few $\times\ 10^9\ M_\odot$) BHs in the SDSS were built by $z = 6$, when the age of the universe is 8×10^8 yr, less than twice the *e*-folding timescale (Haiman & Loeb 2001). As we shall see below, this is not really a problem, since the growth of individual SMBHs at high redshift is dominated by mergers, and not by accretion.

3.2.2. *The Growth of an Individual Black Hole due to Mergers versus Accretion*

A different, but potentially important, ingredient in determining the relative evolution of the SFRD and the BARD at high redshifts is the importance of mergers (see also Volonteri et al. 2003). We next demonstrate that at high redshifts, the buildup of the mass of an individual BH is likely dominated by mergers between BHs. Such mergers may not have any effect on the total quasar emissivity (one can imagine merging all BHs in pairs, resulting in new BHs twice as massive as the original set, while preserving the accretion rate per unit BH mass). However, if mergers are frequent, one can imagine that this may help explain the high-redshift discrepancy between the SFRD and the BARD. For example, one can imagine that a merger event at high redshift delivers new gas and triggers star formation (but may not be able to increase the accretion rate onto BHs, per unit BH mass, if this quantity is already Eddington limited).

The growth rate of BHs due to merging can be obtained from the characteristic dark matter halo number density as follows:

$$\dot{M}_{\rm merg} = M_{\rm nl} \frac{\dot{N}_{\rm nl}}{N_{\rm nl}}. \qquad (14)$$

Here we define the nonlinear dark matter halo mass scale $M_{\rm nl}$ at redshift z as $\sigma(M_{\rm nl}) g(z) = 1$, where $\sigma(M)$ is the rms mass fluctuation in spheres of mass M and $g(z)$ is the growth function at redshift z. For simplicity, we define the space density of halos as $N_{\rm nl} \equiv M_{\rm nl} dN/dM_{\rm nl}$, where dN/dM is the usual (comoving) halo mass function, adopted here from Jenkins et al. (2001), evaluated at $M_{\rm nl}$. Under this assumption, $N_{\rm nl} = N_{\rm nl}(M(t), t)$, and we have the following time derivative:

$$\dot{N}_{\rm nl} = M_{\rm nl} \frac{d^2 N}{dM_{\rm nl}\, dt} + \dot{M}_{\rm nl} \frac{dN}{dM_{\rm nl}}. \qquad (15)$$

The first term on the right-hand side vanishes by definition ($d^2N/dM\, dt \sim 0$ at the nonlinear mass scale). As a result, we find that

$$\dot{M}_{\rm merg} = 0.13 \times 0.1 \times 1.3 \times 10^{-3} M_{\rm nl} \frac{\dot{N}_{\rm nl}}{N_{\rm nl}} \sim 1.7 \times 10^{-5} \dot{M}_{\rm nl}. \qquad (16)$$

This last result reflects the fact that without any accretion, the individual BH masses would grow only by coalescence of the BHs during halo mergers, and therefore the typical BH mass would simply track the nonlinear dark halo mass scale. In order to describe the growth of BHs, rather than that of halos, we have assumed in equation (16) that a fraction 0.13 of the total mass in each halo is baryonic, the mass of stars is 10% of the baryons, and the mass of the central BH is 1.3×10^{-3} that of the stars. We note further that the coalescence rate of galaxies can be slower than that of their dark halos because the dynamical friction time in the common DM halos can be long (Murali et al. 2002; Ghigna et al. 2000; Menci et al. 2003). This delay is increasingly important at lower redshifts, and it does not invalidate our conclusion below that mergers are important at $z > 6$ and can build up the large BH masses by $z = 6$.

We next find the growth of an individual BH due to accretion, using the Madau & Pozzetti (2000) SFRD, as follows:

$$\dot{M}_{\rm acc} = 1.3 \times 10^{-3} \frac{\dot{\rho}_S}{N_{\rm nl}}, \qquad (17)$$

where ρ_S is the comoving SFR density as given by equations (7) and (8) and we take 1.3×10^{-3} for the ratio of the BH mass to the spheroid mass from Kormendy & Gebhardt (2001).

Figure 3 shows the mass growth rates due to merging (*solid curve*) and accretion (*dashed curve*). From this figure, we learn that at redshifts $2 < z < 4$ the growth is dominated by mergers, while at low redshift ($z < 2$) the growth is proportional to the SFR.[9] According to the figure, at very low redshifts ($z < 0.5$) mergers are again important; however, this regime is unphysical because of the so-called overmerging problem in the Press-Schechter formalism: the nonlinear mass scale grows to that corresponding to clusters of galaxies; the galaxies, however, may preserve their identities, and hence it is no longer clear that the growth of BHs by coalescence in merging galaxies tracks this mass scale.

4. RADIATIVE EFFICIENCY AND DUTY CYCLE OF AGNs

In this section we define and derive the radiative efficiency and the duty cycle of AGNs, quantities that served as inputs in the previous sections. In the interest of clarity, let us first

[9] Note that at high redshifts ($z > 4$) the mass buildup by mergers actually exceeds $\sim$10% of the Eddington rate, in accordance with § 3.2.1.

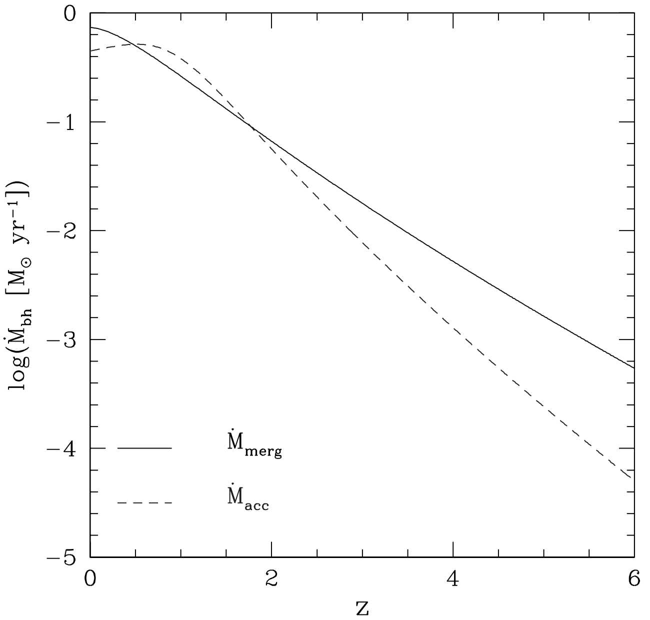

FIG. 3.—Growth rate of an individual BH with the characteristic BH mass at each epoch, from accretion (*dashed curve*) and mergers (*solid curve*), as a function of redshift. Also shown is the growth rate corresponding to accretion that is limited at all times to a fraction 0.07 of the Eddington rate for the characteristic BH mass at each epoch. [*See the electronic edition of the Journal for a color version of this figure.*]

consider a population of N_g identical galaxies over the Hubble time $t_{\rm H}$, each of which today (i.e., at $t = t_{\rm H}$) hosts a spheroid of mass M_S and a BH of mass $M_{\rm BH}$. Let us further assume that during the entire time elapsed from 0 to $t_{\rm H}$, each BH had only two states: it was either "on" or "off." We identify the "on" state as the active quasar phase, and we define the duty cycle f_Q as the fraction of the time each BH spends in the "on" state. At any given time, the number of active quasars is then $N_Q = f_Q N_g$. In the "on" state, the BH grows by accretion at the rate $\dot{M}_{\rm BH}$ and shines at the (bolometric) luminosity L_Q with a radiative efficiency ϵ, defined as the fraction of the rest-mass energy of the infalling gas converted to radiation. The remaining fraction $(1-\epsilon)$ of the rest mass then leads to the growth of the BH mass (Yu & Tremaine 2002). A simple exercise in algebra shows that

$$\frac{\epsilon}{1-\epsilon} = \frac{E_Q^T}{M_{\rm BH}^T c^2} = \frac{f_Q t_{\rm H} L_Q N_g}{M_{\rm BH} N_g c^2} = \frac{t_{\rm H} L_Q N_Q}{M_{\rm BH} N_g c^2}. \quad (18)$$

Here c is the speed of light; the numerator represents the total light emitted by all BHs, and the denominator represents the total mass in BHs today. In the third equality, we have used $N_Q = f_Q N_g$. Note that the last term involves only quantities that are, in principle, directly observable and that it is *independent* of the duty cycle (Soltan 1982). Equation (18) describes the entire galaxy population, but a similar equation applies to individual galaxies: $L_Q f_Q t_{\rm H} = \epsilon M_{\rm BH} c^2$. This last expression does have a dependence on the duty cycle, which can therefore be written as

$$f_Q = \frac{N_Q}{N_g} = \frac{\epsilon M_{\rm BH} c^2}{t_{\rm H} L_Q}. \quad (19)$$

The simple toy model above demonstrates that (1) the radiative efficiency can be obtained independently of the duty cycle and (2) the duty cycle can be obtained two different ways, based on either the *number* or the *characteristic BH mass* of quasars. While the former method is conceptually more straightforward, as we shall see below, the latter avoids the divergence in the number of quasars and galaxies (due to the steep observed slope of the LFs at the faint end).

The step forward to a more realistic situation is to allow a distribution of galaxy and BH masses, as well as corresponding quasar luminosities, and to allow the BH masses and luminosities to change with time. In this case, equation (18) can be straightforwardly generalized to obtain the global average radiative efficiency (Soltan 1982). The total energy output of all quasars over all times per unit volume is given by

$$u_Q = -A_{\rm bol} \int_0^{z_{\rm max}} dz \frac{dt}{dz} \int_{L_{\rm min}}^{L_{\rm max}} dL\, L \Phi_Q(L, z), \quad (20)$$

where $\Phi_Q(L, z)$ is the observed quasar LF, which we take from equation (2) in the B band; $A_{\rm bol}$ is the bolometric correction, which, from the composite quasar spectrum in Elvis et al. (1994), we find to be $A_{\rm bol} = L_{\rm tot}/L_B = 11.2$; and dt/dz is obtained from the time-redshift relation in our chosen ΛCDM cosmology. Note that the integral in equation (20) converges from both above and below in luminosity, as well as in redshift, so that it is insensitive to the integration limits $L_{\rm min}$, $L_{\rm max}$, and $z_{\rm max}$. We find

$$\frac{u_Q}{c^2} = 1.9 \times 10^4\ M_\odot\ {\rm Mpc}^{-3}. \quad (21)$$

The total remnant BH mass in the present universe is given by

$$\rho_{\rm BH} = \int_0^\infty dM\, M \Phi_{\rm BH}(M) = f_{\rm BH} \int_0^\infty dM\, M \Phi_S(M), \quad (22)$$

where $\Phi_{\rm BH}$ and Φ_S are the present-day BH and spheroid mass functions, respectively, and $f_{\rm BH}$ is the BH-to-spheroid mass ratio. We here adopt the spheroid LF from equation (1) and convert it to a spheroid mass function using the M/L ratio for each of the four different types of galaxies in Salucci et al. (1999). Assuming further a constant BH-spheroid mass ratio $f_{\rm BH} = 1.3 \times 10^{-3}$ (Kormendy & Gebhardt 2001), we find

$$\rho_{\rm BH} = 2.49 \times 10^5\ M_\odot\ {\rm Mpc}^{-3}. \quad (23)$$

It is worth noting the contribution to this mass density from the E, S0, Sa/Sab, and Sbc/Scd galaxies separately, which are, respectively, 1.19, 0.63, 0.50, and 0.17 $M_\odot$ Mpc^{-3}. In other words, the bulges of late-type galaxies are not a negligible contribution to the total spheroid mass density (and they dominate the total mass density for low spheroid masses). It is also worth noting that the present-day total spheroid mass density we obtain by integrating the spheroid formation rate in Figure 2 (which was based on a corrected version of the evolution of the total SFR in Madau & Pozzetti 2000) is in good agreement (to within 30%) with the value we find from summing up the inferred local spheroid densities from Salucci et al. (1999; see eq. [1]).

Finally, from the ratio of equations (21) and (23), we find $\epsilon = 0.071$. Our result is in good agreement with recent work by Yu & Tremaine (2002), who find an average efficiency of $\epsilon = 0.077$. It is interesting to note that this agreement holds separately for the QSO light and BH mass density.

TABLE 1

Matching Quasar Light to Black Hole Mass

x	$M_{\rm BH}(x)$	$N^{>}_{\rm BH}(x)$	$L_Q(x,\ 0)/L_{S*}$	$N^{>}_{Q}(x,\ 0)$	$\langle f_{Q,N}\rangle_x$	$\langle f_{Q,M}\rangle_x$
0.1.................	4.33E8	3.92E−5	1.51E0	3.02E−7	0.0077	0.00349
0.2.................	2.67E8	1.14E−4	0.80E0	1.05E−6	0.0092	0.00404
0.3.................	1.82E8	2.29E−4	0.49E0	2.33E−6	0.0102	0.00452
0.4.................	1.27E8	3.94E−4	0.30E0	4.40E−6	0.0112	0.00517
0.5.................	8.91E7	6.28E−4	0.18E0	7.85E−6	0.0125	0.00616
0.6.................	6.03E7	9.69E−4	0.93E−1	1.40E−6	0.0145	0.00784
0.7.................	3.76E7	1.50E−3	0.42E−1	2.67E−6	0.0179	0.0109
0.8.................	2.00E7	2.40E−3	0.14E−1	6.03E−5	0.0251	0.0180
0.9.................	7.05E7	4.48E−3	0.20E−2	2.19E−4	0.0488	0.0433

Note.—The columns show, respectively, the BH mass, the BH space density, quasar luminosity, local QSO space density, and duty cycles (last two columns, computed using two different methods), for galaxies that contain a fixed fraction x (from the first column) of quasar light and both spheroid and BH mass.

Yu & Tremaine (2002) use the recent 2dF quasar LF of Boyle et al. (2000) and a bolometric correction of 11.8, to find $u_Q/c^2 = 2.1 \times 10^4\ M_{\odot}$ Mpc^{-3}. They combine the recent $M_{\rm BH}$-σ relation of Tremaine et al. (2002) with the SDSS velocity function of early-type galaxies in Bernardi et al. (2003a) and make corrections for the additional BH mass in spiral galaxies and for the scatter in the $M_{\rm BH}$-σ relation, yielding $\rho_{\rm BH} = 2.5 \times 10^5\ M_{\odot}$ Mpc^{-3}.

The duty cycle defined in equation (19) is less trivially generalized for a population of evolving galaxies. Nevertheless, if we assume that the duty cycle does not vary with time (but we allow it to be a function of luminosity), then we can still obtain the duty cycle explicitly by comparing the present-day space density of quasars and galaxies. An immediate complication is that both space densities (see eqs. [1] and [2]) diverge at the faint end. We therefore proceed by defining the average duty cycle of all quasars above luminosity $L_Q(x,\ 0)$,

$$\langle f_{Q,N}\rangle_x \equiv \frac{\int_{M_{\rm BH}(x)}^{\infty} dM\,\Phi_{\rm BH}}{\int_{L_Q(x,\,0)}^{\infty} dL\,\Phi_{\rm Q}}, \qquad (24)$$

where $L_Q(x,\ 0)$ is such that QSOs at redshift $z = 0$ brighter than $L_Q(x,\ 0)$ emit a fraction x of the total quasar light $L_Q^T = \int_0^{\infty} dL\,L\,\Phi_Q(L,\ 0)$, and likewise, $M_{\rm BH}(x)$ is such that all BHs more massive than $M_{\rm BH}(x)$ sum up to the same fraction x of the total BH mass at $z = 0$, i.e., to $xM_{\rm BH}^T$.

We may similarly generalize the definition of the duty cycle based on the characteristic BH mass (cf. the last term in eq. [19]), by applying it to an individual present-day BH, as follows:

$$\langle f_{Q,M}\rangle_x \equiv \frac{\epsilon c^2 M_{\rm BH}(x)}{\int_0^{\infty} L_Q(x,t)\,dt}. \qquad (25)$$

In principle, the time integral in the denominator on the right-hand side must be taken over the (typical) luminosity history of the present-day BH of mass $M_{\rm BH}$. If mergers play an important role in the growth of this BH, then the integral must be performed separately and summed over each branch along the "merger tree." In practice, we do not know this merger history, and instead we simply define the time-dependent luminosity $L_Q(x,t)$ such that QSOs at cosmic time t brighter than $L_Q(x,t)$ emit a fraction x of the total quasar light at that epoch, $L_Q^T = \int_0^{\infty} dL\,L\,\Phi_Q(L,t)$. This definition assumes only that a monotonic relation is maintained between $M_{\rm BH}$ and L_Q at all times, and it gives the correct duty cycle in the limit that mergers do not dominate the mass growth of individual BHs (note that the bulk of the total quasar light is emitted in a narrow peak around redshift $z = 2.2$; ~50% of the time integral in eq. [25] is contributed between $1.6 < z < 2.8$). As we showed in the previous section (see Fig. 3), mergers are likely important at $z > 3$ but do not dominate the growth of individual BHs at lower redshifts.

The duty cycles obtained from both methods are listed in Table 1. We find that $\langle f_{Q,N}\rangle_{0.1} \simeq 0.008$ and $\langle f_{Q,N}\rangle_{0.9} \simeq 0.05$. Most importantly, our two definitions above do not guarantee that equations (24) and (25) give the same values. Here we find that the two methods agree well on the high-mass end, while $\langle f_{Q,M}\rangle$ is systematically lower by a factor of ~2 toward the low-mass end. One interpretation of this finding is that the most massive BHs gain nearly all their mass by accretion, while mergers contribute a comparatively larger fraction of the mass of lower mass BHs.

Our results for the duty cycle being at the percent level are in good agreement with theoretical expectations (Ciotti & Ostriker 1997, 2001; Yu & Tremaine 2002) and are also similar to the values derived in Haehnelt et al. (1998), who obtained a *QSO lifetime* of $t_Q \simeq 10^7$ yr at a Hubble epoch of $\simeq 10^9$–10^{10} yr, or, in terms of the duty cycle, $f_Q = t_Q/t_{\rm H} \simeq 10^{-2}$ to 10^{-3}. Similar quasar lifetimes can be independently derived from the spatial clustering of quasars (Haiman & Hui 2001; Martini & Weinberg 2001).

Before we conclude this section, we stress a well-known puzzle presented by the difference in the QSO LFs in the optical and X-ray bands and its consequences for our analysis. We repeated the derivation of the accretion efficiency from the X-ray LF (Miyaji et al. 2001), and we find $\epsilon \sim 0.045$, about 2 times lower than we obtain from the B band. In both cases we applied bolometric corrections from Elvis et al. (1994): $L_{\rm tot}/L_{\rm X} = 38.1$ and $L_{\rm tot}/L_B = 11.2$. If all QSOs emitted intrinsically with the Elvis spectrum, the above two numbers should agree. However, they differ by a factor of 2. What does this mean? The simplest resolution is to assume that QSOs emit a universal spectrum, but the mean value of their flux ratio $L_{\rm X}/L_B$ is a factor of $71/45 = 1.6$ times smaller than for the Elvis et al. (1994) sample (which consists of 47 unobscured QSOs). Indeed, bolometric X-ray corrections by a factor of ~2 higher than that given by Elvis et al. (1994) are commonly adopted in the literature (e.g., Elvis, Risaliti, & Zamorani 2002; Fabian & Iwasawa 1999 and references therein). This implies $L_{\rm X}/L_B \approx 0.18$. However, under the assumption of a universal spectrum, the total number of quasars

$$N(>L_{\min}) = 4\pi \int_0^\infty dz\, \frac{dV}{dz\, d\Omega} \int_{L_{\min}}^\infty dL\, \frac{d\Phi}{dL}(L, z) \quad (26)$$

should then be equal in the optical and in the X-rays, if one uses the appropriate lower limits, $L_{\min,\mathrm{X}} = 0.18 L_{\min,B}$. We find that this leads instead to $N_\mathrm{X} = 2N_B$. In other words, under the assumption of a universal spectrum, the flux ratio L_X/L_B can be derived two ways: either from the total number, or from the total light, of quasars. With the published LFs, these methods give $L_\mathrm{X}/L_B \approx 0.5$ and ≈ 0.2, respectively. This proves that quasars as a population with their measured X-ray and optical LFs cannot have a universal spectrum. One could have obtained this conclusion by directly comparing the X-ray and optical LFs, which have different shapes and redshift evolutions (see, e.g., Ueda et al. 2003; Hasinger et al. 2003; although a better agreement is found between the optical and the X-ray LFs over a wide range of luminosity and redshift by Franceschini et al. 1994). The discrepancy can be resolved if low-luminosity objects are preferentially excluded from optical samples: e.g., a large number of Seyfert galaxies escape from detection in the optical surveys but are detected in the X-rays. Another possible resolution of this puzzle is that the duty cycle in the X-rays is $\sim$3 times longer than in the optical. This would be supported by the recent results of Barger et al. (2001), who found that a large fraction of optical galaxies are *Chandra* AGN sources, implying a long X-ray activity cycle, $\sim$0.5 Gyr.

5. DISCUSSION AND CONCLUSIONS

In this paper we discussed a simple, empirically based model for the growth of SMBHs at the center of spheroidal stellar systems. Motivated by accumulating evidence for the strong link between the formation of spheroids and BHs, we hypothesized the *simplest possible form of this connection*, namely, that star formation in spheroids and BH fueling are proportional to one another, at all cosmic epochs and in all spheroids, regardless of their size.

The main conclusions that arise from this hypothesis (augmented with a few other reasonable assumptions) are as follows. This simple model accounts for the puzzling steep decline of the characteristic luminosity of quasars from redshift $z \approx 2$ to $z = 0$: the residual star formation in spheroidal systems is today limited to a small number of bulges, characterized by stellar velocity dispersions a factor of 2–3 smaller than those of the elliptical galaxies hosting QSOs at $z \gtrsim 2$. We explored a very simple consequence of our hypothesis: the redshift evolutions of the QSO emissivity and of the star formation history in spheroids should be roughly parallel to each other. We find this result to be broadly consistent with the evolution of both the global SFR and the QSO emissivity, both of which exhibit a peak at redshift $z \sim 2$. However, a closer look reveals possibly interesting discrepancies at both low and high redshifts.

At low redshifts, the spheroid formation rate, obtained by making simple corrections to the total SFR, appears to decline by a factor that is $\sim$3 times larger than the decline in QSO emissivity. A possible solution we note to resolve this discrepancy is fueling of quasar BHs at low redshifts by the mass lost in winds from a passively evolving stellar spheroid population, formed at earlier epochs. A tentative discrepancy also exists at high redshifts ($z \gtrsim 2$, beyond the peak of QSO activity), where the evolution of the SFR appears significantly flatter than that of quasar emissivity. While a population of hitherto undetected, obscured AGNs at high redshift (with the obscured fraction increasing toward high z) could resolve this discrepancy, we offered an alternative, physical explanation: quasar fueling rates at high redshift are limited to a fraction $\sim$10% of the Eddington accretion rate. This limit depends linearly on the characteristic BH mass and would therefore imprint a steep evolution of the quasar LF as the characteristic mass scale builds up exponentially. We also note that the masses of individual BHs at high redshift are not limited by accretion (at the Eddington or some modified Eddington rate), since we find that mergers dominate over accretion in determining the growth of objects at epochs $z \gtrsim 2$.

Throughout this paper we have derived conclusions from empirically based hypotheses and avoided a model-dependent interpretation. Nevertheless, it is interesting to place our results in the context of recent works that have addressed the joint evolution of quasars and spheroids using hierarchical galaxy formation models (e.g., Kauffmann & Haehnelt 2000; Volonteri et al. 2003; Granato et al. 2001, 2002, 2004; Wyithe & Loeb 2003). A common feature in such models is that spheroids form predominantly at high redshift, during an epoch when mergers between galactic halos are frequent. This merger activity then triggers both star formation and BH fueling at high redshift. The high-redshift starburst uses up most of the gas reservoir, and the merger rates drop. As a result, the BHs are "starved" and star formation slows down at low redshift. The luminosities of quasars drop substantially below the Eddington value, which then results in a decrease in their characteristic luminosity. As we argued above, this is quite different from our interpretation, namely, that the decline in the characteristic quasar luminosity is due to the decreasing size of the typical quasar host galaxy toward lower redshift. However, interestingly, the apparent difference can be reconciled in models (such as the ones by Granato et al. 2001, 2002, 2004) that include "inverted" hierarchical galaxy formation. There have been suggestions that more massive protogalaxies form their stars *earlier* than their less massive counterparts, despite the collapse of their dark matter halos in the usual hierarchical sequence. The inversion can occur as a result of severe feedback that significantly delays star formation preferentially in low-mass systems (as proposed in Granato et al. 2001, 2002, 2004). We note that a similar effect can arise in the usual hierarchical structure formation theories. At a given redshift, more massive objects are composed of small structures that had formed earlier; these progenitor objects may have formed a significant fraction of stars early on. In this case, at low redshift, the low-mass systems are preferentially still active, in terms of both star formation and BH fueling rate. The latter effect can be identified as a possible physical reason for our finding that low-z quasar hosts are smaller than the high-z hosts. In reality, it is likely that both the decrease in host size and the decrease in the Eddington ratios contribute to the observed decline in the characteristic quasar luminosity.

Finally, given our demographic assumptions, we compute the average duty cycle, i.e., the fraction of time SMBHs spend in the "on" state, as $(3-6) \times 10^{-3}$, depending on BH mass, and we also find the mean bolometric radiative efficiency, $\epsilon = 0.071$, when averaged for the entire SMBH population.

The considerations in this paper are tentative but empirically based. It should soon be possible to considerably tighten constraints on the simple picture of coeval formation of BHs and spheroids: the redshift evolution of the SFR in spheroid systems should be derivable by a more detailed analysis of the SDSS galaxy sample out to at least $z \sim 0.5$.

We thank Günther Hasinger, Martin Haehnelt, Martin Rees, Alvio Renzini, Scott Tremaine, and Qingjuan Yu for useful discussions. We also thank an anonymous referee for a careful reading of the manuscript and for constructive criticism that improved this paper. L. C. was supported by MURST, contract CoFin2000, and grant ASI I/R/105/00. Z. H. was supported by NSF grants AST 03-07200 and AST 03-07291.

REFERENCES

Abraham, R., et al. 1996, ApJ, 471, 694
Andreon, S., & Ettori, S. 1999, ApJ, 516, 647
Bahcall, N. A., Ostriker, J. P., Perlmutter, S., & Steinhardt, P. J. 1999, Science, 284, 1481
Barger, A. J., Cowie, L. L., Bautz, M. W., Brandt, W. N., Garmire, G. P., Hornschmeier, A. E., Ivison, R. J., & Owen, F. N. 2001, AJ, 122, 2177
Bennett, C. L., et al. 2003, ApJS, 148, 1
Benson, A. J., Frenk, C. S., & Sharples, R. M. 2002, ApJ, 574, 104
Bernardi, M., et al. 2003a, AJ, 125, 1817
———. 2003b, AJ, 125, 1849
Binney, J. 1999, in The Radio Galaxy Messier 87, ed. H.-J. Röser & K. Meisenheimer (New York: Springer), 116
Binney, J., & Tabor, G. 1995, MNRAS, 276, 663
Blandford, R. D. 1999, in ASP Conf. Ser. 182, Galaxy Dynamics, ed. D. R. Merritt, M. Valluri, & J. A. Sellwood (San Francisco: ASP), 87
Boyle, B. J, Shanks, T., Croom, S. M., Smith, R. J., Miller, L., Loaring, N., & Heymans, C. 2000, MNRAS, 317, 1014
Bridle, S. L., Lahav, O., Ostriker, J. P., & Steinhardt, P. J. 2003, Science, 299, 1532
Cavaliere, A., Giacconi, R., & Menci, N. 2000, ApJ, 528, 77
Cavaliere, A., & Vittorini, V. 2002, ApJ, 570, 114
Ciotti, L., D'Ercole, A., Pellegrini, S., & Renzini, A. 1991, ApJ, 376, 380
Ciotti, L., Haiman, Z., & Ostriker, J. P. 2001, in The Mass of Galaxies at Low and High Redshift, ed. R. Bender & A. Renzini (Berlin: Springer), 106
Ciotti, L., & Ostriker, J. P. 1997, ApJ, 487, L105
———. 2001, ApJ, 551, 131
———. 2004, in AIP Conf. Proc. 703, Plasmas in the Laboratory and in the Universe: New Insights and New Challenges, ed. G. Bertin, D. Farina, & R. Pozzoli (New York: AIP), in press
Ciotti, L., & van Albada, T. S. 2001, ApJ, 552, L13
Davies, R., Efstathiou, G., Fall, S. M., Illingworth, G., & Schechter, P. L. 1983, ApJ, 266, 41
De Propris, R., Stanford, S. A., Eisenhardt, P., & Dickinson, M. 2003, Ap&SS, 285, 43
Dressler, A., et al. 1987, ApJ, 313, 42
Ellis, R. S., Abraham, R. G., & Dickinson, M. 2001, ApJ, 551, 111
Elvis, M., Risaliti, G., & Zamorani, G. 2002, ApJ, 565, L75
Elvis, M., et al. 1994, ApJS, 95, 1
Faber, S. M., & Jackson, R. E. 1976, ApJ, 204, 668
Fabian, A. C., & Iwasawa, K. 1999, MNRAS, 303, L34
Ferrarese, L., & Merritt, D. 2000, ApJ, 539, L9
Franceschini, A., Hasinger, G., Miyaji, T., & Malquori, D. 1999, MNRAS, 310, L5
Franceschini, A., La Franca, F., Cristiani, S., & Martin-Mirones, J. M. 1994, MNRAS, 269, 683
Fukugita, M., Hogan, C. J., & Peebles, P. J. E. 1998, ApJ, 503, 518
Gebhardt, K., et al. 2000, ApJ, 539, L13
Ghigna, S., Moore, B., Governato, F., Lake, G., Quinn, T., & Stadel, J. 2000, ApJ, 544, 616
Graham, A. W., Erwin, P., Caon, N., & Trujillo, I. 2001, ApJ, 563, L11
Granato, G. L., De Zotti, G., Silva, L., Bressan, A., & Danese, L. 2004, ApJ, 600, 580
Granato, G. L., De Zotti, G., Silva, L., Danese, L., & Magliocchetti, M. 2002, Ap&SS, 281, 497
Granato, G. L., Silva, L., Monaco, P., Panuzzo, P., Salucci, P., De Zotti, G., & Danese, L. 2001, MNRAS, 324, 757
Haehnelt, M. G., Natarajan, P., & Rees, M. J. 1998, MNRAS, 300, 817
Haiman, Z., & Hui, L. 2001, ApJ, 547, 27
Haiman, Z., & Loeb, A. 1998, ApJ, 503, 505
———. 2001, ApJ, 552, 459
Haiman, Z., & Menou, K. 2000, ApJ, 531, 42
Hasinger, G., et al. 2003, in The Emergence of Cosmic Structure, ed. S. S. Holt & C. Reynolds, in press (astro-ph/0302574)
Hogg, D. W., et al. 2002, AJ, 124, 646
Jenkins, A., et al. 2001, MNRAS, 321, 372
Jimenez, R., Verde, L., Treu, T., & Stern, D. 2003, ApJ, 593, 622
Kaspi, S., Smith, P. S., Netzer, H., Maoz, D., Jannuzi, B. T., & Giveon, U. 2000, ApJ, 533, 631
Kauffmann, G., & Haehnelt, M. 2000, MNRAS, 311, 576
King, A. R. 2003, ApJ, 596, L27
Kormendy, J., & Gebhardt, K. 2001, in AIP Conf. Proc. 586, Proc. of the 20th Texas Symposium on Relativistic Astrophysics, ed. J. C. Wheeler & H. Martel (Melville: AIP), 363
Leitherer, C., et al. 1999, ApJS, 123, 3
Madau, P., Haardt, F., & Rees, M. J. 1999, ApJ, 514, 648
Madau, P., & Pozzetti, L. 2000, MNRAS, 312, L9
Madau, P., Pozzetti, L., & Dickinson, M. 1998, ApJ, 498, 106
Magorrian, J., et al. 1998, AJ, 115, 2285
Margoniner, V. E., de Carvalho, R. R., Gal, R. R., & Djorgovski, S. G. 2001, ApJ, 548, L143
Martini, P., & Weinberg, D. H. 2001, ApJ, 547, 12
Menci, N., Cavaliere, A., Fontana, A., Giallongo, E., Poli, F., & Vittorini, V. 2003, ApJ, 587, L63
Metevier, A. J., Romer, A. K., & Ulmer, M. P. 2000, AJ, 119, 1090
Miyaji, T., Hasinger, G., & Schmidt, M. 2001, A&A, 369, 49
Monaco, P., Salucci, P., & Danese, L. 2000, MNRAS, 311, 279
Murali, C., Katz, N., Hernquist, L., Weinberg, D. H., & Davé, R. 2002, ApJ, 571, 1
Newberry, M. V., Kirshner, R. P., & Boroson, T. A. 1988, ApJ, 335, 629
Page, M. J., Stevens, J. A., Mittaz, J. P. D., & Carrera, F. J. 2001, Science, 294, 2516
Pei, Y. C. 1995, ApJ, 438, 623
Percival, W. L., Miller, L., McLure, R. J., & Dunlop, J. S. 2001, MNRAS, 322, 843
Porciani, C., & Madau, P. 2001, ApJ, 548, 522
Proctor, R. N., Sansom, A. E., & Reid, I. N. 2000, MNRAS, 311, 37
Salucci, P., et al. 1999, MNRAS, 307, 637
Sazonov, S. Yu., Ostriker, J. P., & Sunyaev, R. A. 2004, MNRAS, 347, 144
Sheth, R. K., et al. 2003, ApJ, 594, 225
Silk, J., & Rees, M. J. 1998, A&A, 331, L1
Small, T. A., & Blandford, R. D. 1992, MNRAS, 259, 725
Soltan, A. 1982, MNRAS, 200, 115
Spergel, D. N., et al. 2003, ApJS, 148, 175
Stoughton, C., et al. 2002, BAAS, 201, 114.12
Tabor, G., & Binney, J. 1993, MNRAS, 263, 323
Thomas, D., Maraston, C., & Bender, R. 2002, in Reviews in Modern Astronomy, Vol. 15, ed. R. E. Schielicke (New York: Wiley), 219
Tremaine, S., et al. 2002, ApJ, 574, 740
Ueda, Y., Akiyama, M., Ohta, K., & Miyaji, T. 2003, ApJ, 598, 886
van Dokkum, P. G., Franx, M., Fabricant, D., Illingworth, G. D., & Kelson, D. D. 2000, ApJ, 541, 95
Vestergaard, M. 2002, ApJ, 571, 733
———. 2004, ApJ, 601, 676
Volonteri, M., Haardt, F., & Madau, P. 2003, ApJ, 582, 559
Wandel, A., Peterson, B. M., & Malkan, M. A. 1999, ApJ, 526, 579
Woo, J.-H., & Urry, C. M. 2002, ApJ, 579, 530
Wyithe, S., & Loeb, A. 2002, ApJ, 581, 886
———. 2003, ApJ, 595, 614
Yu, Q., & Tremaine, S. 2002, MNRAS, 335, 965